一种邻域自适应半监督局部 Fisher 判别分析算法 *

杜　伟，房立清，齐子元

(军械工程学院 火炮工程系，石家庄 050003)

摘　要：针对利用局部化思想解决多模数据的判别分析问题时，根据经验对局部邻域大小进行全局统一设定无法体现局部几何结构的差异性的不足，提出一种邻域自适应半监督局部 Fisher 判别分析（neighborhood adaptive semi-supervised local Fisher discriminant analysis，NA-SELF）算法。该算法在半监督局部 Fisher 判别分析算法的基础上，结合马氏距离和余弦相似度确定初始近邻数，并根据样本空间概率密度估计调整近邻数。通过人工数据集和 5 组 UCI 标准数据集对该算法的特征降维性能进行验证，并与典型的维数约简算法和采用传统 k 近邻方法的判别分析算法进行比较，实验结果表明该算法具备更高的有效性。

关键词：局部邻域；自适应；半监督局部 Fisher 判别分析；维数约简

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.07.0691

Neighborhood adaptive semi-supervised local Fisher discriminant analysis algorithm

Du Wei, Fang Liqing, Qi Ziyuan

(*Dept. of Artillery Engineering*, *Ordnance Engineering College*, *Sijiazhuang* 050003, *China*)

Abstract: For the discriminant analysis of multimodal data, the idea of localization can hardly reflect the difference of local geometric structure according to the global setting of local neighborhood by experience. Aiming at this problem, this paper proposed a neighborhood adaptive semi-supervised local Fisher discriminant analysis (NA-SELF) algorithm. The new algorithm based on the semi-supervised local Fisher discriminant analysis algorithm, obtained the initial neighborhood by combining the Mahalanobis distance and cosine similarity, and adjusted the number of neighbors according to the probability density estimation of sample space. The performance of feature dimensionality reduction using the algorithm was verified by the synthetic datasets and five UCI standard datasets. Compared with several typical dimensionality reduction algorithms and the discriminant analysis algorithm using the traditional k-nearest neighbor method, the experimental results show that the proposed algorithm has higher effectiveness.

Key Words: local neighborhood; adaptive; semi-supervised local Fisher discriminant analysis; dimensionality reductio

0　引言

随着信息技术的发展，许多研究和应用领域需要处理的数据往往存在维数高、含有大量冗余和混叠信息等问题。为避免陷入“维数灾难”，提高效率并充分挖掘原始数据的本质信息，需要对数据进行有效的维数约简。降维技术作为数据预处理的重要手段，在图像处理、模式识别和计算机视觉等领域得到了广泛应用。

降维算法通过采用线性变换或非线性变换，使数据从高维空间映射到低维空间后尽量保持结构特征信息。降维算法按照样本中是否含有类别标签分为有监督降维和无监督降维[1]。属于有监督降维的线性判别分析（linear discrimination analysis，LDA）[2]和属于无监督降维的主成分分析（principle component analysis，PCA）[3]是典型的线性降维方法。然而，无监督降维方法忽略了类别标签的指导，有监督降维方法需要大量的带标签样本，在工程实践中往往成本过高。半监督降维方法综合利用无标签数据和少量有标签数据，取得了很好的效果[4-6]。为解决线性判别分析在多模数据情况下评价能力差的缺陷，许多学者引入局部化的思想，希望用局部信息来挖掘数据的流形结构，例如局部保持投影（locality preserving projection，LPP）[7]、局部 Fisher 判别分析（local Fisher discrimination analysis，LFDA）[8]和边界 Fisher 判别分析（marginal Fisher analysis，MFA）[9]等。现有的算法在构建邻域时往往是根据经验进行全局统一设定，忽略了数据局部几何结构的差异性，从而影响低维投影向量的

基金项目：河北省自然科学基金资助项目（E2016506003）

作者简介：杜伟（1992-），男，山东临沂人，硕士研究生，主要研究方向为机械设备性能检测与故障诊断（lydwei@163.com）；房立清（1969-），男，河北栾城人，教授，博导，主要研究方向为机电液控制系统与技术；齐子元（1980-），男，山东临清人，副教授，博士，主要研究方向为智能检测与诊断、数字信号处理.

类别可分性。因此，根据数据点之间的距离度量自动确定邻域大小，成为一个值得研究的问题[10]。

基于以上分析，本文提出一种邻域自适应半监督局部 Fisher 判别分析（neighborhood adaptive semi-supervised local Fisher discriminant analysis, NA-SELF）算法。该算法结合距离度量和角度相似性度量构建邻域，并利用样本空间概率密度估计自适应调整近邻数，有效克服了 SELF 算法使用全局统一邻域参数的不足。最后，将运用 NA-SELF 算法得到的低维向量输入支持向量机（support vector machine, SVM）进行识别，验证了 NA-SELF 算法的有效性。

1 半监督局部 Fisher 判别分析

Sugiyama 等人[11]将 LFDA 和 PCA 有效融合，提出一种半监督局部 Fisher 判别分析（semi-supervised local Fisher discriminant analysis，SELF）算法。LFDA 通过描述样本局部信息提高了处理多模态数据的能力，但在有标签样本不足时容易陷入过学习，而 PCA 能够利用无标签样本获取全局分布。SELF 算法结合二者优势，兼具 LFDA 利用类别信息指导降维的能力和 PCA 利用无类别信息获取全局分布的能力。

假设给定样本集共包含 D 维特征，C 个类别，记为 $X=\{x_i\in R^D,(i=1,2,...,n',...,n)\}$，其中有类别标签样本 $x_i(i=1,2,...,n')$，类别标签记为 $l_i\in\{1,2,...,C\}(i=1,2,...,n')$。PCA 的全局散度矩阵定义为

$$S^{(t)}=\frac{1}{2}\sum_{i,j=1}^{n}W_{i,j}^{(t)}(x_i-x_j)(x_i-x_j)^T \tag{1}$$

其中：权值 $W_{i,j}^{(t)}=1/n$。LFDA 的局部类间散度矩阵 $S^{(lb)}$ 和局部类内散度矩阵 $S^{(lw)}$ 可定义为下面的逐对形式[11]：

$$S^{(lb)}=\frac{1}{2}\sum_{i,j=1}^{n'}W_{i,j}^{(lb)}(x_i-x_j)(x_i-x_j)^T \tag{2}$$

$$S^{(lw)}=\frac{1}{2}\sum_{i,j=1}^{n'}W_{i,j}^{(lw)}(x_i-x_j)(x_i-x_j)^T \tag{3}$$

其中：权值矩阵 $W^{(lb)}$ 和 $W^{(lw)}$ 定义为

$$W_{i,j}^{(lb)}=\begin{cases}A_{i,j}(1/n'-1/n'_{l_i}) & \text{if } l_i=l_j\\ 1/n' & \text{if } l_i\neq l_j\end{cases} \tag{4}$$

$$W_{i,j}^{(lw)}=\begin{cases}A_{i,j}(1/n'_{l_i}) & \text{if } l_i=l_j\\ 0 & \text{if } l_i\neq l_j\end{cases} \tag{5}$$

其中：n'_{l_i} 为第 $l_i\in\{1,2,...,C\}(i=1,2,...,n')$ 类样本数；相似矩阵 A 的第 (i,j) 个元素 $A_{i,j}\in[0,1]$ 用于描述两个样本 x_i 和 x_j 之间的相似性，且 $A_{i,j}$ 有高斯相似度、k 近邻相似度和局部尺度相似度等多种定义形式[8]，如

$$A_{i,j}=\exp(-\frac{\|x_i-x_j\|^2}{\sigma_i\sigma_j}) \tag{6}$$

其中：σ_i 为样本点 x_i 的局部尺度，定义为 $\sigma_i=\|x_i-x_i^{(k)}\|$，$x_i^{(k)}$ 为 x_i 的第 k 个最近邻点，文献[8]建议设置全局参数 $k=7$。事实上，式（6）体现了样本点之间的局部近邻关系，能够根据具有相同类别标签的数据对的距离远近对权值进行调整。

由式（1）～（3），定义 SELF 的类间散度矩阵和类内散度矩阵：

$$S^{(b)}=(1-\beta)S^{(lb)}+\beta S^{(t)} \tag{7}$$

$$S^{(w)}=(1-\beta)S^{(lw)}+\beta I_d \tag{8}$$

其中：权系数 $\beta\in[0,1]$，I_d 为标准矩阵。权系数 β 使算法兼具 LFDA 和 PCA 的特性，通过调节其值大小，增加了算法的灵活性。显然，当 $\beta=1$ 时 SELF 等价于 PCA，当 $\beta=0$ 时则等价于 LFDA。寻找最佳的投影转换矩阵 T，即求解如下最大化目标函数问题：

$$T=\underset{T\in R^{d\times r}}{\arg\max}[trT^TS^{(b)}T(T^TS^{(w)}T)^{-1}] \tag{9}$$

式（9）转换矩阵的求解等效于式（10）的广义特征向量求取问题。

$$S^{(b)}\alpha=\lambda S^{(w)}\alpha \tag{10}$$

则转换矩阵 T 由式（10）的前 d 个最大广义特征值对应的广义特征向量 $(\alpha_1,\alpha_2,...,\alpha_d)$ 组成。

2 邻域自适应半监督局部 Fisher 判别分析

传统的近邻数设置方法一般分为 K 近邻法和 ε 近邻法两种。SELF 算法在计算局部尺度相似矩阵 A 时采用 k 近邻法构建邻域，且根据经验设置全局统一参数。然而，实际采集到的样本数据在局部几何结构上往往存在差异性，因此不同的样本在低维映射的过程中所需要的近邻样本集不同，对算法的性能产生的影响也不同。为解决这一问题，本文采用邻域参数自适应调整的方法，在提高算法鲁棒性的同时能提高低维特征的识别效果。

2.1 相似性度量

SELF 算法通过计算样本点与其第 K 个最近邻点的欧式距离来描述局部尺度，但欧式距离只能度量样本间的空间位置，不能体现样本整体的集合结构[12]。而马氏距离不受特征量纲选择的影响[10]，余弦相似度[13]利用矢量夹角的余弦来度量相似性。因此，为充分反映样本间的相似性，文中将余弦相似度和马氏距离相结合，即

$$d_{i,j}=(\frac{1-d_{i,j}^c}{2})\times d_{i,j}^m\ (i,j=1,2,...,n) \tag{11}$$

其中：$d_{i,j}^m$ 和 $d_{i,j}^c$ 分别为样本间的马氏距离和余弦相似度，且 $(1-d_{i,j}^c)/2\in[0,1]$。式（11）融合了样本点间的空间位置和夹角信息，相当于为马氏距离附加了取值范围为 $[0,1]$ 的影响因子，两向量夹角越小则影响因子越小，$d_{i,j}$ 越小。

基于上述融合马氏距离和余弦相似度反映数据分布方面的优势，将 SELF 算法中的相似矩阵元素描述如下：

$$A_{i,j}=\exp(-\frac{d_{i,j}^2}{\sigma_i'\sigma_j'}) \tag{12}$$

其中:将 x_i 及其第 k 个最近邻点 $x_i^{(k)}$ 代入式（11）获得局部尺度 σ_i'。由所有样本的相似系数均值 M_i 确定初始近邻数 k_i，

$M_i=(\sum_{j=1}^{n'}a_{i,j})/n'$，相似系数 $a_{i,j}=\exp(-d_{i,j}^2/\sigma^2)$，$\sigma$ 为所有样本之间距离的均值。若相似系数 $a_{i,j}$ 大于 M_i，则 x_j 是 x_i 的近邻样本。显然，通过该方法得到的每一个样本的近邻数 k_i 可能是不相等的。

2.2　邻域参数自适应调整

在构建邻域时，特征相似的样本分布往往较为密集，而相似性较差的样本分布较为稀疏。若能够根据局部区域样本点的概率密度自适应地调整近邻数 k，则降维得到的低维特征更能反映原始数据的本质结构。Parzen 窗概率密度估计[14]是一种非参数概率密度估计方法，它不需要对概率密度函数形式作出假设，而是由数据自身信息估计出总体概率密度。因此，将 Parzen 窗概率密度估计用于邻域构建，对相似度均值 M_i 确定的初始近邻数 k_i 进行自适应调整。

假设 R^D 是包含数据集 $X=\{x_1,x_2,\ldots,x_N\}$ 的 D 维空间，对于数据点 $x_i(i=1,2,\ldots,N)$，Parzen 窗的概率密度估计式为

$$p(x_i)=\frac{1}{V}\sum_{j=1}^{N}\frac{1}{N}\phi(\frac{d(x_i,x_j)}{h}) \tag{13}$$

其中：$V=h^D$ 为窗体体积，N 为数据集样本个数，h 为窗体宽度，$d(x_i,x_j)$ 为根据式（11）计算的 x_i 与 x_j 的距离，$\phi(x)$ 为窗函数，且满足 $\phi(x)>0$，$\int_x \phi(x)\mathrm{d}x=0$。

窗函数选择平滑性较好的正态窗函数[15]，窗宽 $h=k_i$，k_i 为数据点 x_i 的初始近邻数。则数据点 x_i 的邻域概率密度为

$$p(x_i)=\frac{1}{Nk_i^D}\sum_{x_j\in N_{ki}(x_i)}\frac{1}{\sqrt{2\pi}}\exp(-\frac{d_{i,j}^2}{2k_i^2}) \tag{14}$$

由 $p(x_i)$ 可得到数据集所有样本的平均邻域概率密度 $\bar{p}=[\sum_{i=1}^{N}p(x_i)]/N$，并通过下式调整邻域参数 k_i：

$$k(x_i)=\mathrm{floor}[k_i\frac{p(x_i)}{\bar{p}}] \tag{15}$$

其中：floor 为向下取整函数。

分析式（15）可知，当数据点 x_i 附近数据的概率密度大于平均值时，可自动增大近邻数 $k(x_i)$，使得距离较远的数据对对降维产生较小的作用；反之，则可自动减小近邻数 $k(x_i)$，使距离较近的数据对对降维产生更大的作用，从而保持邻域的局部结构，有利于恢复低维数据集的全局结构信息。

2.3　邻域自适应半监督局部 Fisher 判别分析算法流程

半监督邻域自适应局部 Fisher 判别分析（NA-SELF）算法的具体步骤如下：

输入：D 维空间数据样本集 $X=\{x_i\in R^D,(i=1,2,\ldots,n',\ldots,n)\}$，其中有类别标签样本数为 n'，低维特征空间目标维数 $d(d\le D)$。

输出：投影转换矩阵 T，低维特征向量 Y。

a）根据式（12）计算高维空间数据点间的相似系数 $a_{i,j}$，并由相似系数均值 M_i 得到每个样本的初始近邻数 k_i；

b）由 Parzen 窗概率密度估计计算样本的邻域概率密度 $p(x_i)$，并根据式（15）调整邻域参数 k_i，从而构造相似矩阵 A，代入式（4）（5）得到权值矩阵 $W^{(lb)}$ 和 $W^{(lw)}$，进而得到局部类间散度矩阵 $S^{(lb)}$ 和局部类内散度矩阵 $S^{(lw)}$；

c）根据式（10）求解前 d 个最大广义特征值对应的广义特征向量 $\alpha_1,\alpha_2,\ldots,\alpha_d$，即为投影转换矩阵 T，从而可得低维空间特征 $Y=T^TX$。

2.4　算法时间复杂度分析

文中所提 NA-SELF 算法与原始算法的时间复杂度差异主要体现在 NA-SELF 算法流程的步骤 a）b）中，即计算相似矩阵以及对邻域参数进行自适应调整。假设数据样本的总数为 N，原始特征维数为 D，则由式（11）计算余弦相似度和马氏距离的时间复杂度为 $O(DN^2)$；由式（12）重新计算相似度矩阵时，确定初始近邻数的时间复杂度为 $O(N)$；利用式（14）（15）计算数据点的邻域概率密度和调整邻域参数的时间复杂度均为 $O(N)$。设原始 SELF 算法在整个流程中的时间复杂度为 $O(SELF)$，则经过化简后可得 NA-SELF 算法的时间复杂度为

$$O(NA-SELF)=O(SELF)+O(DN^2) \tag{16}$$

根据式（16）可知，改进算法和原始算法时间复杂度的差异主要与样本总数和原始特征维数有关，样本总数和原始特征维数约多则时间复杂度越大。

3　实验与分析

3.1　人工数据实验

在本实验中，分别利用 PCA、LFDA、SELF 和 NA-SELF 等算法对二类人工数据集进行降维，采用可视化比较实验直观地验证降维算法的性能，算法采用 MATLAB R2013a 实现。在每个人工数据实验中，由二元正态分布随机产生 200 组数据，每组数据包含二类各 100 个无类别标签数据和 10 个有类别标签数据，二类数据分别用圆形和三角形表示，无标签和有标签分别用空心和实心表示。图 1~3 为实验 1~实验 3 中的一组数据及不同算法得到的投影方向，直线表示的是一维的投影空间，分别用不同线型绘出。在每个人工数据实验中将一组作为训练样本，另一组作为测试样本，先对测试样本进行降维得到投影转换矩阵，再使用投影转换矩阵对测试样本进行降维。将低维特征输入支持向量机进行训练识别，共进行 100 次实验，识别率的均值如表 1 所示。设定 SELF 和 NA-SELF 算法中权系数 $\beta=0.5$，SELF 算法中近邻数 $k=7$，SVM 的核函数选用径向基核函数，设置惩罚参数 $C=1$，核函数参数 $g=1$。

图 1 所示的二类数据集各有一个模态，无标签样本均值分别为 $(-4,0)$ 和 $(4,0)$，协方差矩阵为 $[4,0;0,4]$，有标签样本均值为 $(-4,0)$ 和 $(4,-3)$，协方差矩阵为二阶单位阵，显然正确的投影方向为水平方向。实验结果显示，PCA 和 NA-SELF 得到了较好的投影方向；LFDA 受偏下的有标签样本的影响，导致投影方向偏差较大；由于 SELF 同时利用了有标签样本和无标签样本，因此投影方向位于 PCA 和 LFDA 之间。

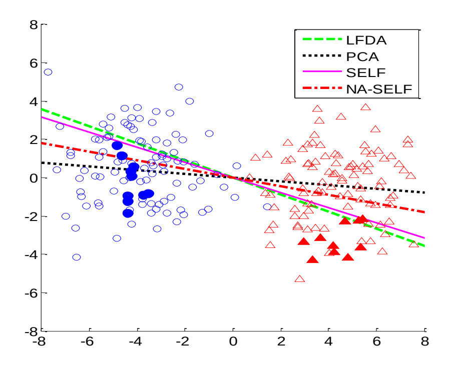


图 1 实验 1 的一组数据

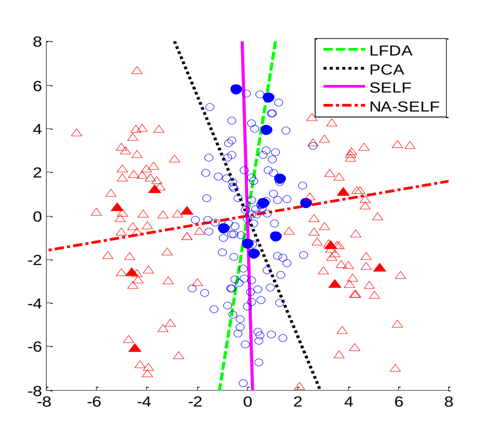


图 2 实验 2 的一组数据

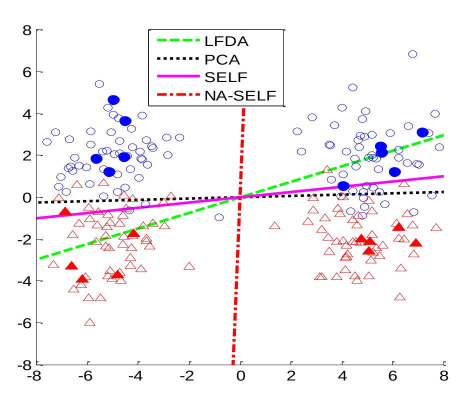


表 1 各种算法的平均识别准确率（%）

算法	实验 1	实验 2	实验 3	平均值
PCA	**96.82**	54.96	54.02	68.60
LFDA	87.98	67.71	62.43	72.71
SELF	92.90	64.46	58.84	72.01
NA-SELF	96.34	**90.44**	**91.51**	**92.76**

图 2 所示的二类数据集分别有一个模态和两个模态，中间一类数据的均值为 $(0,0)$，两侧一类数据的均值分别为 $(-4,0)$ 和 $(4,0)$，无标签样本的协方差矩阵为 $[1,0;0,10]$，有标签样本的均值和协方差矩阵与无标签样本相同，显然正确的投影方向为水平方向。NA-SELF 算法得到了较好的投影方向，而 PCA 选择使数据集方差最大的投影方向，LFDA 选择投影后异类样本的距离平方和较大的方向，因此 PCA 和 LFDA 会选择向垂直方向投影。

图 3 所示的二类数据集各有两个模态，无标签样本均值分别为 $(-8,4)$ $(8,4)$ 和 $(-8,-4)$ $(8,-4)$，协方差矩阵为 $[2,0;0,2]$，有标签样本均值与无标签样本相同，协方差矩阵为二阶单位阵，显然正确的投影方向为垂直方向。由于相同类别两模态样本间的距离大于不同类别相同模态样本间的距离，因此 PCA 会选择水平的投影方向；由于距离较远的同类样本在 LFDA 投影方向的选取中产生较小的作用，因此 LFDA 的投影方向存在一定偏差；NA-SELF 通过自适应调整近邻数，得到的相似矩阵能够更加充分地反映样本数据的局部结构，因此能够得到较好的投影方向。

从表 1 中可以看出，在单模态数据的实验 1 中，NA-SELF 算法的平均识别准确率略低于 PCA，而在具有多模态数据的实验 2 和实验 3 中，NA-SELF 算法比其他算法得到了更高的平均识别准确率，而且 3 个实验结果的平均值也达到最高，表明 NA-SELF 算法具有较为明显的优势，在多模态数据的降维处理上具备更好的适用性。

3.2 UCI 数据实验

从 UCI 机器学习数据库中选取 5 个标准数据集进行维数约简，并将低维特征输入支持向量机进行分类识别。实验中使用的 UCI 数据集如表 2 所示。

表 2 UCI 数据集信息

数据集	类别数	特征维数	训练样本	测试样本
Ionosphere	2	34	100	251
Wine	3	13	95	83
Iris	3	4	60	90
Vehicle	4	18	400	446
Segment	7	18	700	1610

为了便于对比，分别利用 PCA、LFDA、SELF 和 NA-SELF 等算法进行比较实验。其中，SELF 算法中近邻数 $k=7$，SELF 和 NA-SELF 算法的参数 β 采用 5 折交叉验证从 $\{0.1,0.3,0.5,0.7,0.9\}$ 中获得，SVM 的参数设置与 3.1 相同，训练样本中有类别标签样本数与无类别标签样本数按 $1:3$ 随机分配。首先以 Wine 数据集的降维结果为例进行分析。图 4 为利用各种算法将 Wine 数据集降至 5 维时，训练样本低维特征集前 3 个矢量的三维空间分布图。

分析图 4 可知，PCA 的降维效果较差，不同类别的特征集出现了较为严重的混叠；LFDA 仅利用少量有类别标签的样本进行降维，因此各个类别也存在一定程度的混叠；SELF 算法同时利用大量无类别标签样本和少量有类别标签样本，降维后各个类别基本能够分离；NA-SELF 采用马氏距离和余弦相似度相结合的方法能够反映样本点的空间位置和夹角信息，得到的相似性更精确，因此可得到更好的降维效果。

图 5 为各种算法随着选取的降维维数不同，Wine 数据集测试样本的识别准确率。为了比较不同的相似性度量方法对降维效果的影响，将基于欧式距离的 NA-SELF 算法以及基于马氏

距离和余弦相似度相结合的 NA-SELF 算法也进行比较。从图 5 中可以看出，采用不同的维数约简算法和降维维数，数据集的识别准确率均存在差异，而 NA-SELF 算法在一定范围内取得了最高的分类精度。

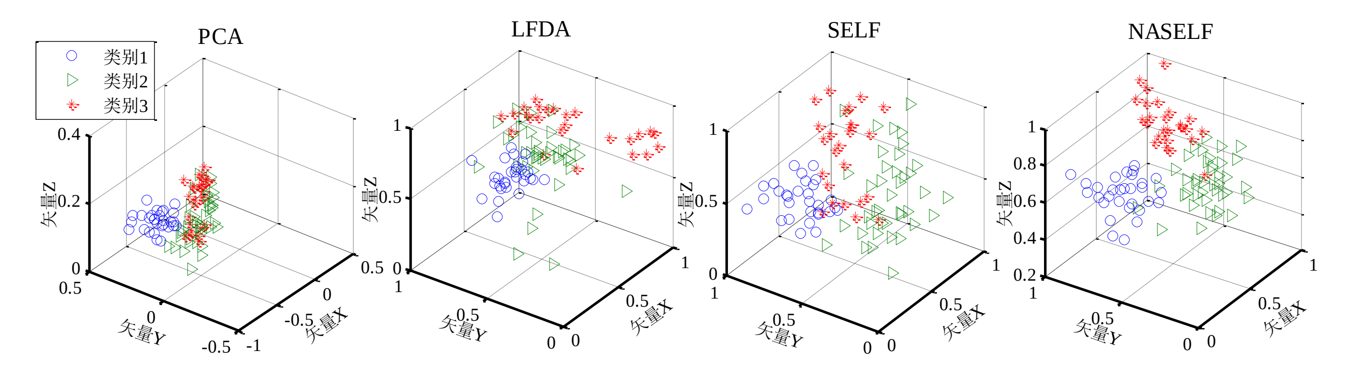


图 4　Wine 数据集各种算法维数约简结果对比

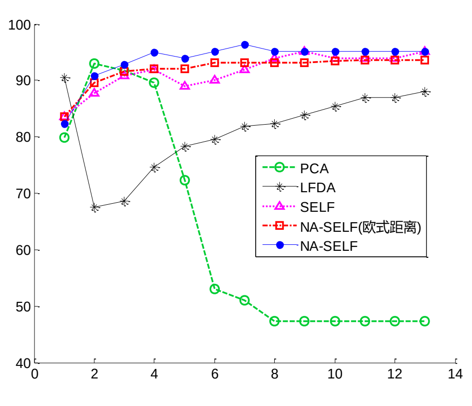


图 5　Wine 数据集测试样本的识别准确率

各种算法在 5 个数据集分别进行 100 次实验，平均识别准确率如表 3 所示。平均识别准确率为每次实验得到的最高识别准确率的平均值，括号中为标准差，同时给出了直接使用原始数据（None）进行分类的平均识别准确率。

根据测试结果可知，由于未经维数约简的原始数据特征集中含有较多的冗余信息，因此大部分数据集降维前的识别率低于降维后的识别率。PCA 具备较好的稳定性，但其属于线性降维方法，忽略了样本数据的非线性结构，而 LFDA 在有类别标签样本不足时可能陷入过学习，因此 PCA 和 LFDA 的识别准确率几乎都低于 SELF 算法。由于 SELF 选取全局统一的邻域参数，所以 SELF 的识别准确率和稳定性相对于 NA-SELF（欧式距离）较低，而文中所提算法采用的相似性度量方法能够更充分反映样本间的相似性，所以识别准确率在 5 个数据集中有 3 个达到最优，且在所有数据集的识别率平均值上也达到了最优。为了比较 NA-SELF 算法与原始 SELF 算法的时间复杂度，表 4 列出了两种算法的平均测试时间，以及改进算法相对于原始算法测试时间增长的百分比。

表 3　各算法的平均识别准确率（%）

算法	Ionosphere	Wine	Iris	Vehicle	Segment	平均值
None	72.11（2.36）	93.39（3.38）	90.33（1.44）	66.95（2.03）	68.82（0.89）	78.32（2.02）
PCA	69.72（1.28）	92.78（2.14）	90.02（1.58）	61.26（1.79）	63.60（2.67）	75.48（1.89）
LFDA	74.13（2.52）	90.83（1.50）	77.78（2.93）	**68.88（2.55）**	69.13（3.14）	77.16（2.53）
SELF	73.92（2.03）	92.93（2.52）	92.56（2.45）	64.45（3.01）	78.70（2.96）	80.51（2.60）
NA-SELF（欧式距离）	**74.81（1.87）**	90.98（2.15）	92.96（2.12）	66.73（2.26）	78.76（2.87）	81.25（2.25）
NA-SELF	74.72（1.95）	**95.16（2.17）**	**95.55（2.25）**	66.95（2.22）	**80.75（2.85）**	**82.63（2.29）**

表 4　测试时间对比　　/ms

耗时	Ionosphere	Wine	Iris	Vehicle	Segment
SELF	203.90	198.83	197.15	228.48	279.47
NA-SELF	269.37	264.58	247.39	318.31	428.79
百分比	32.11%	33.07%	25.48%	39.32%	53.43%

分析表 4 可知，由于改进算法的时间复杂度高于原始算法，因此测试时间略长，并且耗时增长的百分比随着测试样本数量和特征维数的增长而变大，说明数据自身的属性对于算法改进前后的时间复杂度差异具有较大的影响。因此，在模式识别的实际应用中，应充分考虑数据自身的属性，在处理样本数和特征维数相对较少的多模数据时，文中所提方法具有很好的适用性。另外，在对识别准确率要求较高而对计算效率要求次之的场合，也可以将文中所提算法用于多模数据的维数约减。

4　结束语

本文提出了一种邻域自适应半监督局部 Fisher 判别分析算法。该算法采用马氏距离和余弦相似度相结合的方法描述样本

间的相似性，并在构建邻域时利用 Parzen 窗概率密度估计对近邻数进行自适应调整，有效避免了人为选择的随意性，且具有更好的局部几何结构特征表达能力。通过对人工数据集和 5 个 UCI 标准数据集进行维数约简和分类识别的结果表明，相比于典型的 PCA、LFDA 和 SELF 等算法，NA-SELF 算法可得到更优的投影空间和可区分度更高的低维特征向量，基于马氏距离和余弦相似度相结合的相似性度量方法比基于欧氏距离的方法具备更高的有效性。然而在半监督降维算法中，权系数值目前还是利用交叉验证的方法获得，如何快速得到有效的权系数将是本文后续的研究方向之一。

参考文献：

[1] 谢钧, 刘剑. 一种新的局部判别投影方法 [J]. 计算机学报, 2011, 34 (11): 2243-2250.

[2] Zhai L D, Ding Z Y, Jia Y, et al. A word position-related LDA model [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25 (6): 909-925.

[3] Nguyen V H, Golinval J C. Fault detection based on Kernel Principal Component Analysis [J]. Engineering Structures, 2010, 32 (11): 3683-3691.

[4] Yang W Y, Liang W, Xin L, et al. Subspace Semi-supervised Fisher Discriminant Analysis [J]. Zidonghua Xuebao//Acta Automatica Sinica, 2009, 35 (12): 1513-1519.

[5] 苏祖强, 汤宝平, 刘自然, 等. 基于正交半监督局部 Fisher 判别分析的故障诊断 [J]. 机械工程学报, 2014, 50 (18): 7-13.

[6] 杨昔阳, 邓朝阳, 李志伟. 半监督模糊 Fisher 降维分析 [J]. 厦门大学学报: 自然科学版, 2015, 54 (6): 869-875.

[7] Yu J B. Bearing performance degradation assessment using locality preserving projections [J]. Expert Systems with Applications, 2011, 38 (6): 7440-7450.

[8] Sugiyama M. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis [J]. Journal of Machine Learning Research, 2007, 8: 1027-1061.

[9] Yan S, Xu D, Zhang B, et al. Graph Embedding And Extension: A General Framework For Dimensionality Reduction [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2006, 29 (1): 40-51.

[10] 张晓涛, 唐力伟, 王平, 等. 自适应邻域构造流形学习算法及故障降维诊断 [J]. 振动、测试与诊断, 2016, 36 (6): 1210-1215.

[11] Sugiyama M. Semi-supervised local Fisher discriminant analysis for dimensionality reduction [J]. Journal of Machine Learning Research, 2010, 78 (1-2): 35-61.

[12] 刘志勇, 袁媛. 基于测地距离的半监督增强 [J]. 计算机工程与应用, 2011, 47 (21): 202-204.

[13] 李文博, 王大轶, 刘成瑞. 一类非线性系统的故障可诊断性量化评价方法 [J]. 宇航学报, 2015, 36 (4): 455-462.

[14] 杨望灿, 张培林, 吴定海, 等. 基于改进半监督局部保持投影算法的故障诊断 [J]. 中南大学学报: 自然科学版, 2015, 46 (6): 2059-2064.

[15] 刘晗, 张庆, 孟理华, 等. 基于 Parzen 窗估计的设备状态综合报警方法 [J]. 振动与冲击, 2013, 32 (3): 110-114.